Journal de recherche scientifique du Moyen-Orient 23 (6) : 1215-1221, 2015
ISSN1990-9233

DOI : 10.5829/idosi.mejsr.2015.23.06.22271

Reconnaissance d'entités nommées dans les documents d'actualité sur la criminalité à l'aide d'une combinaison de classificateurs

Hafedh Ali Shabat et Nazlia Omar

Centre de technologie de l'IA, FTSM, Université Kebangsaan Malaisie,
UKM, 43000 Bangi Selangor, Malaisie

Résumé : Le volume croissant d'informations sur la criminalité générées et facilement disponibles sur le Web rend
Le processus de récupération, d'analyse et d'utilisation manuelle des informations précieuses contenues dans de tels textes est très difficile.
tâche. Ce travail se concentre sur la conception de modèles pour extraire du Web des informations spécifiques à la criminalité. Ainsi, ceci
L'article propose un cadre d'ensemble pour la tâche de reconnaissance d'entités nommées criminelles. L'objectif principal est de gérer efficacement
intégrer des ensembles de fonctionnalités et des algorithmes de classification pour synthétiser une procédure de classification plus précise.
Premièrement, trois algorithmes de classification de texte bien connus, à savoir Naïve Bayes, Support Vector Machine et
Les classificateurs K-Nearest Neighbor sont utilisés comme classificateurs de base pour chacun des ensembles de fonctionnalités. Deuxièmement, pondéré
La méthode d'ensemble de vote est utilisée pour combiner ces trois classificateurs. Pour évaluer ces modèles, un manuel
un ensemble de données annotées obtenu auprès de BERNAMA est utilisé. Les résultats expérimentaux démontrent que l'utilisation
Le modèle d'ensemble est un moyen efficace de combiner différents ensembles de fonctionnalités et algorithmes de classification pour une meilleure
performances de classement. Le modèle d'ensemble atteint une mesure F globale de 89,48 % pour l'identification du crime
type et 93,36% pour l'extraction d'entités liées à la criminalité. Les résultats du modèle d'ensemble formé avec des
les fonctionnalités surpassent les modèles de base.

Mots clés : Classificateurs du domaine criminel Méthode de combinaison Reconnaissance d'entités nommées Apprentissage automatique

INTRODUCTION

Le volume croissant d'informations sur la criminalité se reproduit. Les documents d'actualité sur la criminalité comprennent
disponible sur le Web, un moyen de récupérer et d'exploiter des détails sur les crimes et ces détails rendent ces documents
des informations pertinentes sont nécessaires pour donner un aperçu des avantages. De nombreuses études ont été menées sur
comportement et réseaux criminels afin de lutter contre la criminalité. performance de ces méthodes basées sur des principes généraux
plus efficacement et plus efficacement. Conception d'un fil de presse électronique. Cependant, les études sont limitées pour
Un système de reconnaissance et d'analyse d'entités nommées criminelles à domaine criminel.
partir de documents d'information en ligne est nécessaire pour aider le Dans cet article, nous présentons la reconnaissance d'entités nommées
autorités à réduire le taux de criminalité. Dans le domaine de la criminalité, système basé sur un cadre d'ensemble pour la criminalité
les analystes de la police et de la criminalité ont besoin d'informations immédiates sur la reconnaissance des entités nommées et l'identification du type de crime
certains cas de crime pour résoudre le crime ou l'empêcher d'accomplir des tâches. Ce système peut reconnaître les types de délits (par exemple, vol,
se reproduise. Les documents d'actualité sur la criminalité comprennent des meurtres, des crimes sexuels, des enlèvements et des drogues) et des extraits
les détails sur les crimes et ces détails font de ces documents des entités (par exemple, les armes du crime, les lieux et la nationalité)
bénéfique. Les modèles de reconnaissance d'entités nommées sont extraits de documents criminels. L'objectif principal de l'utilisation d'ensemble
ces informations utiles plus rapidement et avec un cadre élevé sont de synthétiser une classification plus précise
précision fiable [1]. procédure. Premièrement, trois classifications de textes bien connues
Identification du type de crime et algorithmes d'entité nommée, à savoir Naïve Bayes, Support Vector Machine
la reconnaissance sont des tâches importantes d'extraction d'informations et les classificateurs K-Nearest Neighbour sont utilisés comme
qui traite de la reconnaissance et de la classification des classificateurs de base pour chacun des ensembles de fonctionnalités. Deuxième,
des documents ou des jetons ou des séquences de jetons liés à une méthode d'ensemble de vote pondéré sont utilisés pour combiner
classe ou entité particulière. Dans le domaine criminel, la police et ces trois classificateurs.
les analystes de la criminalité ont besoin d'informations immédiates sur certains
cas de crime pour résoudre le crime ou l'empêcher de se produire

Auteur correspondant : Hafedh Ali Shabat, Centre for AI Technology, FTSM, University Kebangsaan Malaysia, UKM,
43000 Bangi Selangor, Malaisie.

Travaux connexes : Plusieurs modèles NE populaires emploient diverses techniques pour l'extraction de NE dans les articles de journaux criminels de trois pays, à savoir, New domaine. [2] ont développé un modèle utilisant les neurones de la Zélande, de l'Australie et de l'Inde. Le modèle utilise le NER pour réseaux pour acquérir des informations utiles auprès de personnes non structurées déterminer si la peine inclut une scène de crime documents et rapports criminels. Les informations récupérées sont la localisation. L'approche employée est la conditionnelle introduit dans une base de données comme étape ultérieure pour un champ aléatoire qui est une méthode d'apprentissage automatique utilisée pour d'autres modèles d'extraction de données et de textes pour identifier, vérifier la présence d'informations sur l'emplacement de la scène du crime modèles liés à la criminalité. Les informations extraites sont dans une phrase. forme structurée et est essentiel pour les systèmes d'exploration de données. L'objectif de cette étude est de développer une procédure [3]. Afin d'identifier les modèles de criminalité et d'accélérer l'extraction de données à partir de documents d'actualité sur la criminalité en ligne processus de résolution de crimes, [4] a utilisé un algorithme de clustering. par l'approche de combinaison de votes en fusionnant les Suite aux mesures de mise à niveau, la machine à vecteurs de support k-means (SVM), les baies Naïve (NB) et une procédure de regroupement a été utilisée pour renforcer les classificateurs moyens du k-voisin le plus proche (KNN). Cette procédure peut pour déterminer les modèles de criminalité. De véritables forces de l'ordre soient employées pour identifier les types de délits les informations provenant du bureau du shérif ont été utilisées pour le (vol, meurtre, crime sexuel, enlèvement et drogue) et le application de cette procédure.

[5] ont créé un système IE d'extraction d'informations adapté au domaine criminel. Ce système est capable d'obtenir des informations sur la criminalité à partir de rapports de police, de rapports de témoins et de documents d'actualité. Il récupère informations sur les personnes, les armes, les véhicules, les lieux, le temps et des vêtements. Une évaluation de ce système a été menée dans le cadre d'apprentissage automatique pour les textes criminels automatiques avec l'utilisation de deux formats différents de classification de texte et du système criminel NER. La méthodologie documents, nommément rapports de police et rapports de témoins. se compose de deux tâches principales : l'identification du type de crime et Ces rapports ont été rassemblés à partir de forums, de blogs et d'informations sur la criminalité NER sites Internet d'agences.

Un prototype a été développé pour l'identification des types de crimes dans le contexte arabe [6]. Deux techniques ont été appliquées pour les processus de reconnaissance : la première technique dépendait entièrement de l'identification directe à l'aide de répertoires géographiques et la seconde technique est un modèle basé sur des règles dans lequel les règles sont construites sur la base d'une liste d'indicateurs de criminalité qui comprend divers mots-clés pertinents [7]. Développement d'une procédure comparable qui reconnaît d'autres informations connexes en dehors du type de crime. Ces informations proposent un cadre dans ce travail. comprend la nationalité de la victime et le lieu de résidence la scène du crime. Cette procédure comprend un indicateur pour Description de la ressource linguistique : Utilisation d'un gérer la langue arabe. la technique d'apprentissage automatique dépend de la

[8] Création d'une technique qui repose sur la disponibilité naturelle de données d'entraînement annotées. De telles données sont procédures de traitement du langage et emploie celles généralement créées manuellement par des humains ou des experts dans le domaine. Modèle inférentiel sémantique. Le système dans leur domaine pertinent. Pour concevoir un crime NER, la présente étude développé est personnalisé pour l'utilisation collaborative d'un ensemble de données annotées à partir de documents criminels. environnements sur Internet. Cette méthode, appelée Wiki. Chaque mot du corpus de formation était étiqueté comme étant un crime. Crimes, est utilisé pour acquérir deux armes criminelles fondamentales, le lieu du crime et les entités de nationalité. Les données les entités des pages Web en ligne, à savoir les scènes de crime utilisées dans cette recherche, ont été collectées auprès du Malaisien lieu et type de crime. [9] L'agence nationale de presse (BERNAMA) a proposé un modèle IE. Type de crime, met l'accent uniquement sur l'extraction d'informations liées aux armes, au lieu du crime et à la nationalité impliquée au vol qui comprend l'emplacement de la scène du crime, c'est-à-dire ont été annotés et classés manuellement.

son adresse. Les informations sur le vol sont extraites de

obtention d'informations à partir de documents criminels concernant les armes utilisées pour commettre un crime, l'emplacement de la scène du crime, les nationalités des personnes impliqué, etc.

Conception de la recherche : Cette étude présente un ensemble

Le processus commence par des procédures de prétraitement pour éliminez les données imparfaites, bruitées (dénuées de sens) et sporadiques. En tout état de cause, le prétraitement des données est une opération opérationnelle exigence préalable à l'exécution d'autres opérations d'exploration de données procédures. Déterminer les termes perceptifs supérieurs pour la formation et les tests, plusieurs extractions de fonctionnalités Des techniques ont été mises en œuvre. Enfin, un certain nombre de les procédures de catégorisation de l'apprentissage automatique sont utilisées pour la reconnaissance des entités nommées et l'identification des types de crimes. La figure 1 illustre l'architecture du

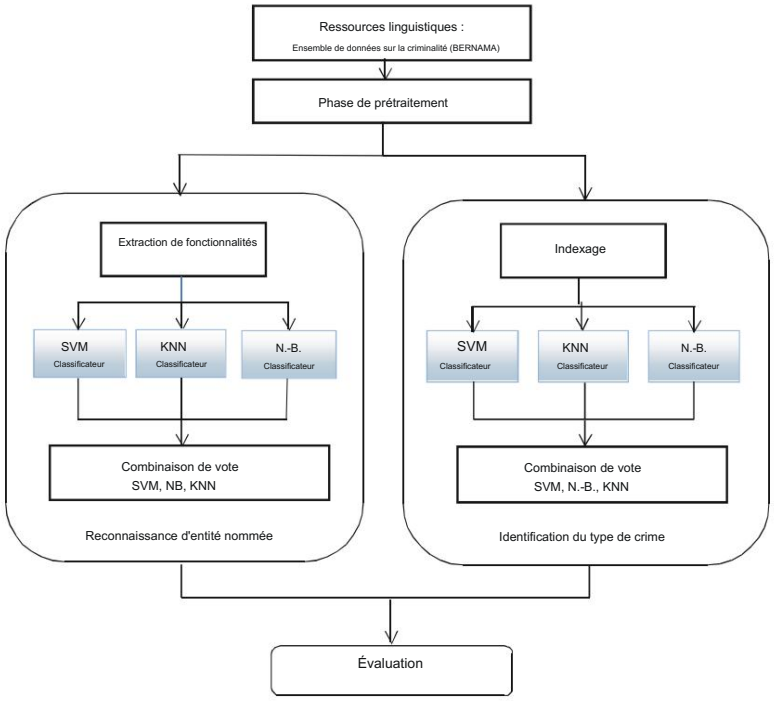

Fig. 1 : Architecture du framework proposé

Tableau 1 : Échantillon de texte criminel annoté avec des balises POS

Mot	Mot d'étiquette	Mot d'étiquette	Mot d'étiquette	Étiqueter
UN	DT par	DANS	Bandar NNP ce soir RB	
Sécurité NN trois CD nus			NNP Selangor NNP	
Garde NN armé JJ		Tonne	NNP a dit	VBD
Les hommes du VBD NNS Hussein NNP étaient-ils les				DT
Shot VBN boutique NN activé			PNN	…… ……
Mort	JJ dans	DANS	ici RB	…… …….

Pré-traitement : les données ayant été collectées auprès de journaux malaisiens et de sites de médias sociaux, elles incluaient généralement des données bruyantes. Par conséquent, le prétraitement des données est crucial à l'aide d'approches d'apprentissage automatique.
Avant qu'un NER et un type de crime puissent être identifiés, chaque document criminel doit passer par les étapes de prétraitement. Dans ce système, pendant la phase de pré-traitement, l'identification du type de crime nécessite la tokenisation, la suppression et la radicalisation des mots vides et NER nécessite la tokenisation et une partie du discours (POS) dans la phase de pré-traitement. Le tableau 1 montre un échantillon de texte criminel après la phase de prétraitement.

Extraction de fonctionnalités : dans toutes les procédures de classification, l'extraction de fonctionnalités est cruciale car elle améliore les performances des tâches de classification en termes de rapidité et d'efficacité d'apprentissage. L'objectif de l'extraction de caractéristiques est la conversion de chaque mot en un vecteur de valeurs de caractéristiques.

Un ensemble de fonctionnalités a été défini pour l'acquisition de données provenant de sources en ligne concernant les nationalités, les armes et les lieux de crime. Par la suite, le regroupement de ces fonctionnalités sous les trois ensembles de fonctionnalités principaux de (a) fonctionnalités établies sur le marquage POS, (b) fonctionnalités établies sur les affixes de mots et (c) fonctionnalités établies sur le contexte est effectué. Ces ensembles de fonctionnalités sont également utilisés pour la représentation des mots dans le corpus.
Le tableau 2 présente un résumé de ces ensembles de fonctionnalités utilisés respectivement pour l'extraction de l'arme, de l'emplacement et de la nationalité.

Indexation : Le type de crime est identifié dans ce travail grâce à la classification des documents criminels. Par conséquent, le document est converti d'une version texte intégral en une version vecteur de document pour rendre le document plus simple et plus facile à traiter. Représentation de documents : méthode utilisée pour réduire la complexité des documents et les rendre plus faciles à manipuler. Ce processus est accompli en convertissant l'édition textuelle complète du document en un vecteur de document. Le modèle spatial vectoriel (VSM) est sans doute la représentation de document la plus fréquemment utilisée [10]. La classification de texte présente un problème qui attribue automatiquement les documents criminels non étiquetés à des types de crimes prédéfinis. Dans le cadre de la classification des types de délits, la représentation textuelle transforme le contenu des documents texturaux en un format compact, de sorte que

Tableau 2 : résumé des ensembles de fonctionnalités

Catégorie de fonctionnalités	Nom de la fonctionnalité	Fonctionnalité
Affixes de mots	F1	Préfixe1
	F2	Préfixe2
	F3	Préfixe3
	F4	Suffixe1
	F5	Suffixe2
	F6	Suffixe3
Fonctionnalités basées sur le contexte	F7	Mot précédent (taille de la fenêtre 2)
	F8	Mot suivant (taille de la fenêtre2)
	F9	Nombre de mots indicateurs d'armes avant (taille de la fenêtre 7)
	F10	Nombre de mots indicateurs d'armes après (taille de la fenêtre 7)
	F11	Distance en mots entre le mot actuel et les mots indicateurs avant le mot actuel.
	F12	Distance en mots entre le mot actuel et les mots indicateurs après le mot actuel.
Basé sur le point de vente	F13	La partie du discours du mot est-elle un nom

les documents peuvent être reconnus et classés par un classificateur [11]. Dans le VSM, un document est représenté comme un vecteur dans le terme espaces = (w ,w$_1$,….,$_2$w), où |V| est la taille du vocabulaire. La valeur de wi représente comment le terme wi contribue à la sémantique du document D. Classification des types de crimes sous forme de texte

$$\bar{\alpha} = \sum_{j=1}^{n} \sum_{i=1}^{n} \alpha_j \alpha_i a_j a_i x_j x_i \text{ , argmin}_{\alpha} \quad (1)$$

$$\text{Sujet de } : \sum_{i=1}^{n} \alpha_i = 0 \text{ ; } 0 \leq \alpha_i \leq C \quad (2)$$

La tâche de classification emprunte le terme traditionnel de pondération Naive Bayes (NB) : la procédure NB est fréquemment schémas du domaine de la recherche d'informations, tels que ceux utilisés pour classer les avis. Étant donné un vecteur de caractéristiques TF.IDF [12]. Dans la présente étude, en identifiant le type de table, l'algorithme détermine la possibilité de reculer dans le crime est réalisé en classifiant le texte. lequel l'examen est lié à une gamme de classes et le l'algorithme l'attribue à la classe avec le plus grand arrière

Algorithmes de classification : la majorité du potentiel de la machine. La procédure NB, qui correspond à un les méthodologies d'apprentissage comportent deux étapes. Au cours du modèle stochastique de fabrication de documents, utilise le étape initiale, une formation est dispensée pour la génération d'une règle de Bayes. Aux fins de classer la classe c* avec machine entraînée, alors que l'étape suivante implique le potentiel le plus élevé pour un nouveau document d, le classification. Une évaluation du calcul d'apprentissage automatique sélectionné est la suivante :

méthodologies ont été menées au cours de cette étude. Cependant, pour l'acquisition d'informations sur C = argmaxc P(C | d). crime qui inclut les nationalités (3)

impliquées, le armes utilisées et emplacements des scènes de crime grâce à des documents criminels disponibles en ligne, cette étude s'est basée sur les classificateurs d'apprentissage automatique suivants :

Vous trouverez ci-dessous le calcul du classificateur NB pour la probabilité postérieure.

$$P(c_i \mid d) = \frac{P(c_i) P(d \mid c_i)}{P(d)} \quad (4)$$

Machine à vecteurs de support (SVM) : Cette machine innovante La procédure d'apprentissage a été recommandée par [13]. Conformément à K-Nearest Neighbour (KNN) : la fonction principale de la conviction de réduire les menaces structurelles liées à cet algorithme d'apprentissage supervisé est la catégorisation des

Dans les conceptions de l'apprentissage informatique, une décision fait apparaître les données en fonction de leur ressemblance avec les données prédéfinies. Avec utilisé par le SVM bifurque les points de données de formation vers l'utilisation d'une variété de mesures de distance. prendre des décisions qui sont vérifiées par le classificateur, évaluer la ressemblance entre un objet non classé vecteurs de support. Ces vecteurs supports sont reconnus comme objet de données et les données prédéfinies. Par la suite, le composants actifs dans l'ensemble de formation. Pendant que l'algorithme calcule l'écart entre les données non classifiées la génération d'un certain nombre de variantes du SVM a été un objet et les k objets les plus proches situés dans le prédéfini enregistré [14], cette enquête a choisi de se concentrer uniquement sur l'ensemble de données de formation. La majorité de classe des k les plus proches sur SVM linéaire. Le choix de cette technique est attribué aux voisins qui constituent la classe déterminée pour le à sa réputation de classification de textes de bonne qualité [15]. objets de données non classifiés [16]. La distance euclidienne est L'optimisation SVM (double forme) s'exprime ainsi : fréquemment utilisée pour la mesure de distance :

$$D_{Euclidien}(x,y) = \sqrt{\sum_{i=1}^{m}(x_i - y_i)^2} \quad (5)$$

où x = (x , attributs x_2 , ..., x_m) et (y , y_2 ,..., y_m) signifie le m de deux échantillons.

Combinaison de vote : les algorithmes de vote prennent le classificateur de sorties ; de certains classificateurs comme entrée et sélectionnez une classe qui a et la quatrième expérience a utilisé la combinaison de vote été sélectionné par la plupart des classificateurs comme sortie. méthode. Toutes ces expériences ont été appliquées pour identifier La règle de vote compte les prédictions des cinq types de crimes (vol, meurtre, crimes sexuels, enlèvement). classificateurs puis attribue l'échantillon de test x à la classe i avec et médicaments) en classant les documents dans le corpus, les prédictions les plus composantes. en fonction du contenu des documents. Dans le NER,

$$\hat{O}_{J.} = \sum_{k=1}^{D} I(\arg\max O_{kj}) \quad (6)$$

où I(•) est la fonction indicatrice. La règle de somme combine les sorties des composants à l'aide de l'équation suivante :

$$\hat{O}_{J.} = \sum_{k=1}^{D} \hat{O}_{kj}, \quad (7)$$

ce qui équivaut à la moyenne des productions sur classificateurs (règle moyenne).

RÉSULTATS ET DISCUSSION

Évaluer les modèles proposés, les données des crimes, des enlèvements et des drogues. La classe de vol contenait les corpus ont été collectés à partir de 383 documents du Malaysian National News, représentant 27,31 % du corpus ; le Agence (BERNAMA). Les types de crimes, armes, meurtres contenaient 298 documents, représentant localisation du crime et nationalité contenues dans ces 21,25% du corpus ; la catégorie des crimes sexuels contenait 299 les documents étaient annotés et classés manuellement. les documents, représentant 21,32% du corpus ; le Dans cette étude utilisant le processus de validation croisée, la classe de kidnapping contenait 200 documents, représentant le corpus a été divisé au hasard en 10, soit 14,26 % du corpus ; et la classe de médicaments contenait 222 sous-échantillons. Un seul sous-échantillon a été retenu comme documents, représentant 15,83 % du corpus. Quatre les données de validation pour tester le modèle et les expériences restantes sont appliquées : la première expérience évalue le 9 sous-échantillons ont été utilisés comme données de formation. Le classificateur NB, la deuxième expérience évalue le SVM le processus de validation croisée a ensuite été répété 10 fois par le classificateur, la troisième expérience évalue le KNN (10 plis). L'ensemble d'apprentissage représente les valeurs d'entrée du classificateur et la quatrième expérience évalue le classificateur le modèle de classification de NB, SVM et KNN. La méthode de combinaison. Le tableau 3 présente un résumé des Le corpus représente les entrées de données dans ce modèle. résultats expérimentaux utilisant le NB, le SVM et le KNN

Les évaluations standard de précision, de rappel et de mesure F classificateurs, ainsi que l'algorithme de vote, pour identifier ont été utilisées pour évaluer l'efficacité des capacités de reconnaissance le type de crime. d'entités nommées disponibles dans le modèle proposé. La précision (P), le rappel, la mesure F et la macro-moyenne (F1) sont les principaux critères d'évaluation de l'efficacité des systèmes NER criminels [17].

Cadre expérimental : dans cette étude, un certain nombre de et KNN sont appliqués à l'ensemble de l'espace de fonctionnalités. Après des expériences ont été menées pour évaluer les performances évaluées par la méthode de combinaison de classificateurs.

des modèles proposés. Ces expériences ont été menée pour identifier le type de crime et reconnaître les entités nommées associées dans les documents criminels.

Pour l'identification du type de crime, quatre expériences ont été menées. La première expérience a été réalisée en utilisant le Classificateur du Nouveau-Brunswick ; la deuxième expérience a utilisé le SVM

quatre expériences ont également été menées. La première l'expérience a été réalisée à l'aide du classificateur NB ; le la deuxième expérience a utilisé le classificateur SVM ; le troisième l'expérience a appliqué le classificateur KNN ; et le quatrième L'expérience a utilisé la méthode de combinaison de votes. Tous ceux-ci des expériences ont été appliquées pour identifier les entités (arme, nationalité et lieu du crime) du crime documents. Ces expériences ont appliqué un ensemble de fonctionnalités qui comprennent trois types : affixes de mots, basés sur le contexte fonctionnalités et basées sur les points de vente.

Expériences pour l'identification du type de crime : dans le crime module d'identification de type, un corpus de formation composé de 1 402 documents criminels ont été utilisés. Les données du corpus ont été divisés en cinq catégories de crimes : vol, meurtre, sexe

Expériences de classification d'entité nommée : dans ce expérience, la performance globale de chaque individu classificateur et classificateurs combinés dans une entité criminelle l'extraction a été examinée. Les trois classificateurs, NB, SVM et KNN sont appliqués à l'ensemble de l'espace de fonctionnalités. Après

Tableau 3 : performances pour chaque type de délit

	N.-B. (%)	MVS (%)	KNN (%)	Algorithme de vote (%)
Vol	69,5	83.2	82,0	87,4
Meurtre	61,6	88,9	72,9	90,1
Crimes sexuels	68,7	87.1	83,3	88,6
Enlèvement	64,4	81.2	84,7	87,4
Drogues	69.1	89,8	87,8	93,9
Macro-F-Mesure	66,7	86.04	82.15	89.48

Tableau 4 : performances pour chaque type d'entité

	N.-B. (%)	MVS (%)	KNN (%)	Algorithme de vote (%)
Armes	86,73	91.08	82.35	93,3
Nationalité	94.02	96.25	84.23	97,5
Emplacement	87,66	89.28	81,62	89.28
Macro-F-Mesure	89.47	92.20	82,73	93.36

Le tableau 4 présente un résumé des résultats expérimentaux utilisant les classificateurs NB, SVM et KNN, ainsi que les algorithme de vote, pour extraire la nationalité, l'arme et le lieu du crime.

DISCUSSION

Selon les expériences d'identification du type de crime, le résultat le plus élevé obtenu par les classificateurs individuels a été celui du classificateur SVM avec une précision de 86,04 % et le résultat le plus bas a été obtenu par le classificateur NB avec une précision de 66,6 %. La méthode de combinaison de vote a donné un résultat d'exactitude de 89,48 %, soit supérieur à celui de tous les classificateurs individuels. En outre, et couverts (arme, nationalité et lieux du crime), le résultat le plus élevé obtenu par les classificateurs individuels a été celui du classificateur SVM avec une précision de 92,2 % et le résultat le plus bas a été obtenu par le classificateur KNN avec une précision de 82,73 %. La méthode de vote combinée a abouti à une précision de 93,36 %, ce qui était supérieur à celui de tous les classificateurs individuels.

CONCLUSION

En évaluant l'ensemble des résultats obtenus à chaque fois en appliquant un classificateur, la précision la plus élevée était de 89,48 % pour l'identification des types de délits et de 93,36 % pour l'identification des entités obtenues en utilisant la méthode combinée. Le résultat a montré que le modèle proposé était important pour identifier le type de crime et extraire les entités nommées associées à partir des documents criminels. La recherche les résultats ont été comparés à ceux des classificateurs individuels et ont confirmé la plus grande précision.

LES RÉFÉRENCES

1. Kumar, N. et P. Bhattacharya, 2006. Reconnaissance d'entité nommée en hindi à l'aide de memm. les actes du rapport technique, IIT Bombay, Inde.
2. Chau, M., JJ Xu et H. Chen, 2002. Extraire des entités significatives de la police rapports narratifs. Dans les actes de l'édition 2002 conférence nationale sur la recherche sur le gouvernement numérique, Société de gouvernement numérique d'Amérique du Nord., pp : 1-5.
3. Hauck, RV, H. Atabakhsb, P. Ongvasith, H. Gupta et H. Chen, 2002. Utilisation de Coplink pour analyser les données criminelles selon les expériences du crime nommé entité justice. Ordinateur., 35 : 30-37.
4. Nath, SV, 2006. Détection des modèles de criminalité à l'aide de données exploitation minière. InWeb Intelligence et Agent Intelligent Ateliers technologiques, 2006, WI-IAT 2006 Ateliers, 2006 IEEE/WIC/ACM International Conférence sur, pp : 41-44.
5. Ku, CH, A. Iriberri et G. Leroy, 2008. Crime extraction d'informations auprès de la police et des témoins rapports narratifs. Dans Technologies pour la patrie Sécurité, Conférence IEEE 2008 sur., pp : 193-198.
6. Alruily, M., A. Ayesh et H. Zedan, 2009. Type de crime classification des documents à partir du corpus arabe. Dans Développements en ingénierie des systèmes électroniques (DESE), Deuxième conférence internationale sur., IEEE, pp : 153-159.
7. Alruily, M., A. Ayesh et H. Zedan, 2009. Type de crime classification des documents à partir du corpus arabe. Dans Développements en ingénierie des systèmes électroniques (DESE), Deuxième conférence internationale sur., IEEE, pp : 153-159.

8. Pinheiro, V., V. Furtado, T. Pequeno et D. Nogueira, 2010. Traitement du langage naturel basé sur l'inférence sémantique pour extraire des informations sur la criminalité à partir de texte. En Informatique de Renseignement et de Sécurité (ISI), Conférence internationale IEEE 2010 sur., pp : 19-24.

9. Arulanandam, R., BTR Savarimuthu et MA Purvis, 2014. Extraction d'informations sur la criminalité à partir d'articles de journaux en ligne. Dans Actes de la deuxième conférence Web australasienne., 155 : 31-38. Société australienne d'informatique, Inc.

10. Aas, K. et L. Eikvil, 1999. Catégorisation du texte : A Enquête. ISBN82-539-0425-8.

11. Ko, Y., 2012. Une étude des systèmes de pondération des termes utilisant les informations de classe pour la classification des textes. Dans Actes de la 35e conférence internationale ACM SIGIR sur la recherche et le développement en recherche d'information, pp : 1029-1030.

12. Salton, G. et C. Buckley, 1988. Approches de pondération des termes dans la récupération automatique de texte. Information Traitement et gestion : une revue internationale Traiter, gérer., pp : 513-523.

13. Cortes, C. et V. Vapnik, 1995. Vecteur de support réseaux, Machine Learning, 20 : 273-297.

14. Joachims, T., 1998. Catégorisation de texte avec support les machines vectorielles : Apprentissage avec de nombreux caractéristiques, Springer Berlin Heidelberg, pp: 137-142.

15. Yang, Y. et X. Liu, 1999. Un réexamen du texte méthodes de catégorisation. Dans les actes du 22 Conférence internationale annuelle ACM SIGIR de l'ACM sur la recherche et le développement en matière d'information Récupération., pp: 42-49.

16. Inyaem, U., P. Meesad et C. Haruechaiyasak, 2009. Techniques d'entité nommée pour un événement terroriste extraction et classement. En langage naturel Traitement, 2009, SNLP'09, Huitième Internationale Colloque sur,. IEEE, pages : 175-179.

17. Manning, DC, P. Raghavan et H. Schütze, 2008. Introduction à la recherche d'informations. Cambridge Presse universitaire. ISBN978-0-521-86571-5.